知识管理论坛 ISSN 2095-5472 CN11-6036/C
*Knowledge Management Forum*

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

# 国际语料库语言学研究热点与前沿的信息可视化分析

◎杨柳
上海外国语大学英语学院 上海 200083

**摘要**：[ **目的 / 意义** ] 本研究旨在更好地把握近几年国际语料库研究发展的整体脉络和研究成果，厘清该领域研究的热点问题，探索其研究的前沿。[ **方法 / 过程** ] 以 Web of Science 核心数据库所收录国际期刊在 2005-2017 年间所刊载的研究性论文作为数据来源，借助 BICOMB、Ucinet6 和 CiteSpace 软件对语料库相关研究数据从文献数量与年代分布、来源期刊、研究主体、国家和地区分布、研究机构、被引文献等方面进行计量和知识图谱分析。[ **结果 / 结论** ] 通过分析发现，国际语料库语言学的研究对象涉及到更多语种和文类，并越来越呈现出跨学科、多角度的特点。持续关注的研究热点包括搭配和词典编撰，新热点包括专门用途语言、学术英语、性别、身份、隐喻及话语分析，与构式语法、认知语言学等的结合是研究前沿。

**关键词**：语料库 可视化分析 BICOMB Ucinet6 CiteSpace
**分类号**：H083

**引用格式**：杨柳 . 国际语料库语言学研究热点与前沿的信息可视化分析 [J/OL]. 知识管理论坛 , 2018, 3(4): 208-224[ 引用日期 ]. http://www.kmf.ac.cn/p/142/.

“语料库”来自拉丁语“corpus”，意为“汇总”“文集”。一般认为 1967 年美国布朗语料库的建立和相关论文的发表标志着语料库研究在现代语言学意义上的开端。但是 20 世纪 60 年代的美国盛行理性主义，语料库语言学最初是在欧洲得到发展；英国成为语料库研究的重镇，并形成赞成和反对语料标注两种态度，前者代表如 R. Quirk[1]、G. Leech[2] 和 T. McEnery[3]，后者代表为 J. M. Sinclair[4]。伦敦大学的 R. Quirk 在 1959 年宣布建立“英语用法调查”（The Survey of English Usage）语料库；英国新弗斯学派代表人物 J. M.Sinclair 主持 COBUILD 项目，建成科林斯英语语料库（The Bank of English）；M. Baker 将语料库引入翻译研究 [5]。自此，语料库广泛应用于词典编撰、语法描述、二语习得、文学研究及翻译研究等领域 [6-7]。美国第一次全国性语料库研讨会于 1999 年举办，开始迎头赶上；2001 年第一届语料库语言学国际会议于英国兰卡斯特大学召开，国际交流进一步加强。相对于西方，我国的语料库语言学研究起步较晚，但成果也颇为

作者简介：杨柳（ORCID：0000-0002-3588-0787），副教授，博士，E-mail：youngwillow@126.com。
收稿日期：2018-05-23 发表日期：2018-08-14 本文责任编辑：刘远颖

知识管理论坛，2018(4):208-224
DOI：10.13266/j.issn.2095-5472.2018.021


丰富，最早始于80年代上海交通大学科技英语计算机语料库（JDEST）的建立，2000 年以后相关论文发表数量开始成倍增长。2003 年，首个中国学习者英语语料库建成[8]；2006 年，王克非首次提出“语料库翻译学”的概念[9]；2009 年，首届全国语料库翻译学研讨会在上海交通大学召开；2011 年，首届中国语料库语言学大会在北京外国语大学举行。

目前，随着计算机及网络技术的革新，语料库规模更大，美国杨百翰大学的 iWeb 语料库达到百亿词级；应用软件更优更新；语料库的应用领域更广；文献发表数量与日俱增。为了全面了解近年来国际语料库研究的发展态势，把握该领域研究的热点和前沿问题，本文运用 BICOMB 和 CiteSpace 等工具软件，对发表在 Web of Science 上的 2005-2017 年间国际语料库研究文献进行分析，绘制可视化知识图谱，期待为国内语料库研究提供参考。

## 1 数据来源与研究方法

### 1.1 数据来源

本研究采集的数据来源于 Web of Science（WOS）核心合集，该合集包括 Sciences Citation Index（SCI）、Social Sciences Citation Index（SSCI） 和 Arts & Humanities Citation Index（AHCI）数据库，包括 2005 年至今科学、社会科学、艺术和人文科学领域的世界一流学术性期刊、书籍和会议录。以“corpus” or “corpora”为检索主题词进行检索，文献类型为论文（Article），学科领域限定为语言学（Linguistics 和 Language Linguistics），语种为英语（English），不限定出版时间，截至 2017 年 12 月 5 日共检索到英文文献 5 096 篇，基本涵盖了 2005年以来国际学界语料库研究的重要成果。文献数据包含全文本与引用的参考文献。

### 1.2 研究方法

科学计量可视化软件的优势是迅速处理海量数据，并以可视化方式呈现，直观揭示数据特征。本文根据研究对象和问题，选取了 3 个软件工具，分别是 BICOMB[10]、Citespace 和 Ucinet6。通过 BICOMB 进行核心期刊与期刊共被引分析、高产第一作者分析和高被引作者分析；通过 Ucinet6 进行作者合作分析；通过 CiteSpace 对文献的数量趋势、期刊来源、国家和地区分布、研究机构、共被引文献和前沿热点进行分析。基于 3 个软件对 5 096 篇文献生成的可视化谱图及阐释，呈现出国际语料库语言学在过去 13 年间整体的发展趋势和特点，为后续研究提供参考。

## 2 数据分析和讨论

### 2.1 国际语料库文献数量

文献数量的变化情况是衡量该领域研究进展的重要指标，经统计在 WOS 数据库共收录期刊文献 5 096 篇，年均文献量为 392 篇，各年代文献数量分布如图 1 所示。国际语料库研究从 2005 年开始，该领域的整体研究呈现上升趋势，其趋势可以分为 4 个阶段：①快速发展阶段。2005-2009 年，这一阶段是语料库研究的快速增长时期，文献数量保持稳步递增。②平稳发展阶段。2010-2012 年，此阶段语料库研究论文增长幅度不大，基本都在 400 篇左右。③再提速阶段。2013-2015 年，这一阶段语料库研究又呈现快速增长趋势，并且在 2015 年达到语料库研究文献量的最大值 622 篇。④递减阶段。2015 年至今，这一阶段文献数量逐步递减。

### 2.2 核心期刊与期刊共被引分析

通过对语料库领域相关期刊的分布情况进行研究，有利于了解该领域发文期刊的空间分布，并发现该领域的主流期刊及发展动态；同时，也有助于相关学者了解该领域的研究进展及发文情况。2005-2017 年刊出语料库的 5 096 篇文章分布在 251 个来源出版物，利用 BICOMB 软件对发文期刊进行统计筛选，根据布拉德福定律确定语料库研究文献的核心期刊。将全部文献划分为经典的 3 个区间，对各个区间的文献数和期刊数进行统计得到区域分析表，如表 1 所示：

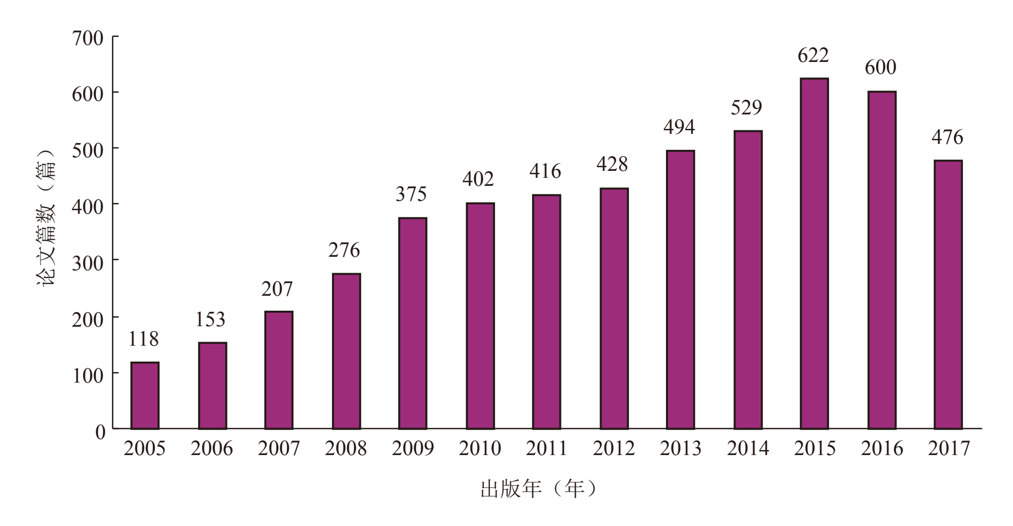



图 1 Web of Science 核心收录的语料库研究文献发表情况

表 1 原始数据集区域分布表

| 区域 | 期刊数（种） | 占期刊总数比例（%） | 载文量（篇） | 占论文总数比例（%） | 平均载文密度（篇/种） | 布拉福德常数（n） |
|---|---|---|---|---|---|---|
| 核心区 | 19 | 7.57 | 1 731 | 33.97 | 91.11 | |
| 相关区 | 44 | 17.53 | 1 670 | 32.77 | 37.95 | 2.315 |
| 外围区 | 188 | 74.90 | 1 695 | 33.26 | 9.0 | 2.07 |
| 合计 | 251 | 100 | 5 096 | 100 | 20.30 | |

按照布拉德福定律，以论文为统计单元，将所有期刊按相关论文数降序排列并划分为论文总数大致相等的 3 个区域，统计各区域的期刊数，判断是否符合 1 ∶ n ∶ $n^2$。此处，1 ∶ n ∶ $n^2$ 可以作为判断分布是否符合布拉德福定律的评估指标；比例越是接近 1 ∶ n ∶ $n^2$，则布拉德福定律越显著，布布拉德福定律的应用也越准确 [11]。表 2 中语料库 3 个区域的期刊数为 19 ∶ 44 ∶ 188，即 1 ∶ 2.315 ∶ 4.27（$2.07^2$），比例系数基本一致，原始数据能较好地满足布拉德福定律描述的条件。根据“核心区 - 相关区 - 外围区”的划分理论，得到 19 种核心期刊，如表 2 所示：

期刊共被引（Cited Journal）分析是指两本或多本期刊被同一篇文献引用的现象，期刊共被引所反映的是各类期刊及学科的关联性，通过期刊共被引分析可以获得某个研究领域的知识基础分布。采用 CiteSpace 软件对上述转化后的数据进行期刊共被引分析。时间分段（Time Slicing）选择 2005-2017 年，时间切片（Years per lice）选择 1，节点类型（Node Types）选择被引期刊（Cited Journal），切片上限（Top N per slice）选择 50，运用软件进行可视化分析结果如图 2 所示，图中节点较大的期刊是在国际语料库研究领域具有较高影响力的期刊。按照中心度（取中心度前 30 位的期刊）排序统计，语料库研究的重要被引期刊见表 3。

通过 CiteSpace 得到节点数 124，连线数 315 的期刊共被引图谱（见图 2）。被引期刊中心度排名前 30 的期刊见表 3。19 种核心期刊见表 2。这些期刊发表的语料库语言学成果最多，影响力最大，应该重点关注。19 种核心期刊中英国出版 5 种，荷兰 4 种，德国 4 种，法国、西班牙、美国、加拿大、智利、南非各 1 种；语

知识管理论坛，2018( 4 ):208-224
DOI：10.13266/j.issn.2095-5472.2018.021



料库语言学刊物 2 种，计算机语言学 3 种，专门用途语言 3 种，词典编撰 1 种，翻译 1 种，语言学 4 种，认知语言学 1 种，语用学 2 种，其他 2 种。发文量最大的前 5 种期刊分别是《语用学期刊》《语料库语言学国际期刊》《专门用途英语》《语料库语言学和语言学理论》和《META：译者期刊》，其中前两本期刊的中心度即影响力也是最高的。从核心期刊与期刊共被引情况可以看到，除自语料库研究发端就与其密切相连的词典编撰、翻译、语法描述等领域外，语用学、专门用途英语、认知语言学也广泛地和语料库语言学产生联系，拓展了研究深度和广度。美国语言学协会的刊物《语言》和德国德古意特出版社出版《认知语言学》发表的论文数量不是最多的，但被引中心度分别为第一和第五，也是语料库语言学的重要参考文献，并且证明了认知语言学和语料库的结合是一个新的研究热点。

表 2 语料库研究核心期刊

| 序号 | 来源期刊 | 发文量（篇） | 百分比(%) | 累计百分比（%） |
|---|---|---|---|---|
| 1 | *JOURNAL OF PRAGMATICS*（语用学期刊） | 295 | 5.788 9 | 5.788 9 |
| 2 | *INTERNATIONAL JOURNAL OF CORPUS LINGUISTICS*（语料库语言学国际期刊） | 156 | 3.061 2 | 8.850 1 |
| 3 | *ENGLISH FOR SPECIFIC PURPOSES*（专门用途英语） | 114 | 2.237 0 | 11.087 1 |
| 4 | *CORPUS LINGUISTICS AND LINGUISTIC THEORY*（语料库语言学和语言学理论） | 96 | 1.883 8 | 12.971 0 |
| 5 | *META*（META: 译者期刊） | 84 | 1.648 4 | 14.619 3 |
| 6 | *ENGLISH LANGUAGE & LINGUISTICS*（英语语言和语言学） | 83 | 1.628 7 | 16.248 0 |
| 7 | *JOURNAL OF ENGLISH FOR ACADEMIC PURPOSES*（学术英语期刊） | 82 | 1.609 1 | 17.857 1 |
| 8 | *COMPUTATIONAL LINGUISTICS*（计算语言学） | 81 | 1.589 5 | 19.446 6 |
| 9 | *NATURAL LANGUAGE ENGINEERING*（自然语言工程） | 79 | 1.550 2 | 20.996 9 |
| 10 | *REVISTA SIGNOS*（符号学期刊） | 78 | 1.530 6 | 22.527 5 |
| 11 | *TEXT & TALK*（文本和谈话） | 75 | 1.471 7 | 23.999 2 |
| 12 | *LEXIKOS*（词典学） | 72 | 1.412 9 | 25.412 1 |
| 13 | *LINGUISTICS*（语言学） | 70 | 1.373 6 | 26.785 7 |
| 14 | *LANGUAGE SCIENCES*（语言科学） | 65 | 1.275 5 | 28.061 2 |
| 15 | *IBERICA*（伊比利亚） | 63 | 1.236 3 | 29.297 5 |
| 16 | *COGNITIVE LINGUISTICS*（认知语言学） | 61 | 1.197 0 | 30.494 5 |
| 17 | *LITERARY AND LINGUISTIC COMPUTING*（文学与语言计算） | 59 | 1.157 8 | 31.652 3 |
| 18 | *LINGUA*（LINGUA：普通语言学国际评论） | 59 | 1.157 8 | 32.810 0 |
| 19 | *LANGUE FRANCAISE*（法语） | 59 | 1.157 8 | 33.967 8 |

图 2 语料库研究领域期刊共被引图谱

表 3 被引期刊中心度排名前 30 的期刊列表

| 中心度 | 被引期刊 | 被引频次 | 中心度 | 被引期刊 | 被引频次 |
|---|---|---|---|---|---|
| 0.46 | *LANGUAGE* | 1 283 | 0.07 | *INTRO FUNCTIONAL GRA* | 295 |
| 0.39 | *INT J CORPUS LINGUIS* | 698 | 0.07 | *LANG SPEECH* | 174 |
| 0.34 | *J PRAGMATICS* | 1 134 | 0.06 | *COMPREHENSIVE GRAMMA* | 380 |
| 0.32 | *LONGMAN GRAMMAR SPOK* | 509 | 0.06 | *J LINGUIST* | 371 |
| 0.31 | *COGN LINGUIST* | 485 | 0.06 | *PRAGMATICS* | 234 |
| 0.21 | *APPL LINGUIST* | 825 | 0.06 | *COMPUTATIONAL LINGUISTICS* | 171 |
| 0.21 | *COGNITION* | 426 | 0.06 | *INT J LEXICOGR* | 83 |
| 0.2 | *J MEM LANG* | 417 | 0.06 | *WOMEN FIRE DANGEROUS* | 76 |
| 0.2 | *TEXT* | 399 | 0.05 | *TESOL QUART* | 452 |
| 0.1 | *LANG VAR CHANGE* | 365 | 0.05 | *J ENGL LINGUIST* | 254 |
| 0.1 | *CAMBRIDGE GRAMMAR EN* | 265 | 0.05 | *LANG COGNITIVE PROC* | 164 |
| 0.1 | *COGNITIVE SCI* | 248 | 0.05 | *DISCOURSE PROCESS* | 161 |
| 0.08 | *ENGL LANG LINGUIST* | 236 | 0.05 | *SPEAKING INTENTION A* | 12 |
| 0.07 | *ENGL SPECIF PURP* | 532 | 0.04 | *CORPUS CONCORDANCE C* | 194 |
| 0.07 | *LINGUIST INQ* | 359 | 0.04 | *J PHONETICS* | 37 |

## 2.3 研究主体分析

研究主体指从事语料库研究的学者，本文从高产第一作者、高被引作者和作者合作三方面分析研究主体。

### 2.3.1 高产第一作者分析

学术影响的广度和深度主要取决于学者所发表的研究成果，通过确定某领域研究的核心作者，可以大致发现该领域的知识地

知识管理论坛，2018( 4 ):208–224
DOI：10.13266/j.issn.2095–5472.2018.021

图，从而促进这一领域的学术交流与合作。通过 BICOMB2.0 软件对文献发文作者情况统计分析，5 096 篇文献共涉及第一作者 3 755 人。根据洛特卡定律，当发文量为 1 篇的作者数占作者总数的比例低于 60% 时，会形成核心作者群 [12]。经统计，2005-2017 年发文量为 1 篇的作者有 2 968 位，约占作者总数的 79.04%，高于洛特卡定律提出的 60% 标准，说明国际语料库领域未能够形成核心作者群。根据普赖斯定律 M=0.749（Nmax）1/2[13]，发文量大于等于 3 的作者为高产第一作者，共 296 人，本文统计发文量为 6 篇及以上的作者，具体如表 4 所示：

表 4 2005–2017 年语料库研究部分作者统计

| 序号 | 作者 | 发文量（篇） | 序号 | 作者 | 发文量（篇） |
|---|---|---|---|---|---|
| 1 | S. T. Gries | 16 | 19 | I. M. P Martinez | 6 |
| 2 | G. M. de Schryver | 12 | 20 | S. Wulff | 6 |
| 3 | K. Hyland | 11 | 21 | J. Flowerdew | 6 |
| 4 | G. Parodi | 10 | 22 | R. Venegas | 6 |
| 5 | D. Biber | 10 | 23 | J. Parkinson | 6 |
| 6 | D. L. Liu | 9 | 24 | L. Flowerdew | 6 |
| 7 | P. Collins | 9 | 25 | E. Taljard | 6 |
| 8 | J. L. B. Arroyo | 9 | 26 | M. Charles | 6 |
| 9 | M. A. Jimenez-Crespo | 8 | 27 | P. Durrant | 6 |
| 10 | D. J. Prinsloo | 8 | 28 | S. F. Chung | 6 |
| 11 | N. C. Ellis | 8 | 29 | L. De Cuypere | 6 |
| 12 | S. A. Crossley | 8 | 30 | K. O'Halloran | 6 |
| 13 | P. Baker | 7 | 31 | L. Anderwald | 6 |
| 14 | M. Hilpert | 7 | 32 | A. Adel | 6 |
| 15 | A. Partington | 7 | 33 | C. Y. Lin | 6 |
| 16 | R. Moon | 7 | 34 | J. Owens | 6 |
| 17 | S.Crossley | 6 | 35 | M. Bednarek | 6 |
| 18 | C. Ruhlemann | 6 | | | |

表 4 显示，国际从事语料库研究的主要学者有 S. T. Gries、G. M. de Schryver、K. Hyland、G. Parodi 和 D. Biber 等人，这几位高产第一作者发表了 10 篇以上的高质量论文，他们是国际语料库研究的领军人物。以 S. T. Gries 等为代表的核心作者总计发文 1 146 篇，约占论文总数的 22.5%，虽未达到普赖斯提出的 50% 标准 [14]，但贡献比较可观。这一方面说明这些核心作者是语料库研究领域的主体，为语料库的发展做出了重要贡献；另一方面还说明语料库研究的学者群学术影响力还不够大，致使核心作者群尚未形成。

为了进一步了解第一作者之间的合作情况，利用 Citespace 对收集的文献进行作者合作分析，得到图 3 所示的作者合作聚类图谱，图中节点代表被引作者，节点越大表示作者的发文量越大。

图 3 中共有 377 个节点，102 条连线，网络密度为 0.001 4。其中，节点的大小与作者发文数量有关，节点间的连线表示作者间的合作关系。从图 3 可知，国际语料库领域形成了以 S. T. Gries、G. M. de Schryver 和 K. Hyland 等为代表的高发文作者群，这些作者是国际语料库领域的开拓者和集大成者。







图 3 语料库研究领域作者知识图谱

下面仅对发文量前 3 位高产作者进行简要分析：S. T. Gries 在语料库研究中累计发表英文论文 16 篇，居于首位。该学者是美国加利福尼亚大学语言学教授、定量语料库语言学家，同时还是一个以认知为导向的使用型语言学家，使用各种不同的统计方法来研究语言的话题，比如使用语料库来研究其主体的同质性与比较、联结与分散测量、N-gram 识别与探索以及其他定量方法。G. M. de Schryver 发表英文论文 12 篇，居于第二位，是比利时根特大学语言和文化系非洲语言学研究教授，主要研究语料库语言学、计算语言学，他和 D. Joffe 一起搭建了 20 个非洲语言语料库。K. Hyland 发表英文论文 11 篇，居于第三位，是英国东安格利亚大学教授，香港大学应用语言学首席教授、应用英语研究中心主任，国际著名应用语言学家，学术写作与语料库分析领域世界领军学者。

2.3.2 高被引作者分析

被引频次是衡量研究成果价值的重要指标，被引频次的高低可以反映出作者在相关领域的影响力，对语料库研究期刊的高被引作者进行分析，可以发现影响语料库研究的重要人物。通过 BICOMB2.0 软件对高被引作者进行统计分析发现，143 400 篇被引文献共涉及作者 85 996 人，以被引频次 100 为节点，得到高被引作者共 43 人，具体情况如表 5 所示：

知识管理论坛，2018( 4 ):208–224
DOI：10.13266/j.issn.2095–5472.2018.021



表 5 2005–2017 年语料库研究高被引作者统计

| 序号 | 被引作者 | 被引次数 | 中心度 | 序号 | 被引作者 | 被引次数 | 中心度 |
|---|---|---|---|---|---|---|---|
| 1 | D. Biber | 925 | 0.39 | 23 | P. J. Hopper | 201 | 0.01 |
| 2 | M. A. K. Halliday | 582 | 0.16 | 24 | M. Davies | 201 | 0 |
| 3 | J. Sinclair | 454 | 0.09 | 25 | W. Chafe | 196 | 0.12 |
| 4 | R. Quirk | 438 | 0.1 | 26 | E. A. Schegloff | 188 | 0.03 |
| 5 | K. Hyland | 418 | 0.23 | 27 | B. Macwhinney | 182 | 0.02 |
| 6 | W. Labov | 373 | 0.13 | 28 | K. Aijmer | 173 | 0.05 |
| 7 | M. Scott | 371 | 0.05 | 29 | J. L. Bybee | 165 | 0.02 |
| 8 | G. Leech | 359 | 0.05 | 30 | E. Goffman | 160 | 0.06 |
| 9 | G. Lakoff | 358 | 0.06 | 31 | D. Bolinger | 152 | 0.09 |
| 10 | R. W. Langacker | 337 | 0.1 | 32 | R. D. Huddleston | 151 | 0.08 |
| 11 | J. Bybee | 330 | 0.2 | 33 | T. Mcenery | 144 | 0.08 |
| 12 | S. Hunston | 289 | 0.11 | 34 | M. Baker | 139 | 0.03 |
| 13 | J. M. Swales | 281 | 0.06 | 35 | S. C. Levinson | 133 | 0.01 |
| 14 | P. Brown | 264 | 0.04 | 36 | D. Crystal | 132 | 0.01 |
| 15 | W. Croft | 250 | 0.03 | 37 | H. H. Clark | 126 | 0.05 |
| 16 | E. C. Traugott | 231 | 0.15 | 38 | H. Sacks | 123 | 0.04 |
| 17 | A. Goldberg | 230 | 0.09 | 39 | N. C. Ellis | 117 | 0.12 |
| 18 | N. Chomsky | 221 | 0.01 | 40 | N. Fairclough | 114 | 0.03 |
| 29 | T. Givon | 217 | 0.03 | 41 | M. Haspelmath | 113 | 0.03 |
| 20 | S. T. Gries | 209 | 0.13 | 42 | P. Baker | 105 | 0.01 |
| 21 | S. Granger | 209 | 0.06 | 43 | A. Wierzbicka | 102 | 0.02 |
| 22 | M. Stubbs | 206 | 0.04 | | | | |

期刊的质量与引文作者密切相关，利用 CiteSpace 对收集的 143 400 篇参考文献进行作者共被引分析，得到节点数 66，连线数 158 的作者共被引图谱，如图 4 所示。每一个节点代表一位被引作者，节点大小表示该作者的被引频次，节点越大表示该作者的被引频次越高。

结合表 5 和图 4 发现，被引频次和中心度排名都在前 20 的作者中，D. Biber、K. Hyland、J. Bybee、M. A. K. Halliday、E. C. Traugott、W. Labov 和 S. T. Gries 都排在前列，这些作者在国际语料库领域均做出了卓越贡献。N. C. Ellis 虽共被引只有 117 次，但中心度为 0.12，表示其研究内容是一个重要的转折点，他将语料库运用到二语习得研究，为其他学者带来了重要启示。

2.3.3 作者合作分析

作者合作水平根据合作密度值来判断。密度指的是网络中各个成员之间联系的紧密程度，是指行动者之间实际联结的数目与他们之间可能存在的最大联结数目的比值，其高低代表群体成员平均互动程度的强弱，密度值越大，成员之间的联系就越密切 [15]。将处理好的矩阵导入 Ucinet6，依次点击 Network-Cohension-Density，进行合著网络的密度分析，可

以得出国际语料库研究主体合作网络整体网络密 Density（matrix average）为 0.0131，标准差 Standard deviation 为 0.243 8，这表明国际语料库研究主体合作水平不高，作者之间的联系较为松散。说明语料库领域研究者团队之间沟通少，如果不同的研究团队之间加强交流，则能给不同的团队注入新的活力，有利于知识的分享和传播，进而促进该领域的发展。

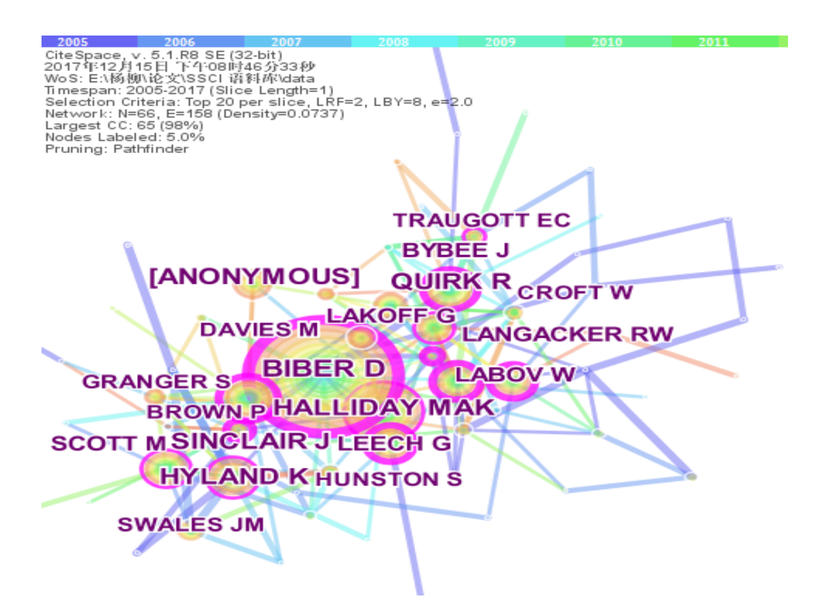



图 4 国际语料库研究高被引作者聚类图谱

同时，合作研究中通常使用合作率（collaboration rate, CR）和合作水平（collaboration level，CL）两个指标度量合作程度。合作率是指作者数大于等 2 的论文数占全部论文数的比例，合作水平一般是用所有论文的平均作者合作度表示 [16]。由此可以得到语料库研究的合作度为 0.82，合作率为 22.36%，合作度和合作率都很低。对独著者去重后分析独著者为 3 755 人，重复人数占近 39.4%，说明作者之间的合作有待加强。

## 2.4 文献国家和地区分布

考察文献的国家和地区是指文献第一作者所在的国家和地区。经统计共有 57 个国家 / 地区对语料库进行了研究，发文量超过 10 篇以上的国家 / 地区见表 6。由表 6 可知，美国、英国、西班牙、德国、比利时、中国和法国等是开展语料库研究的主要国家，说明这些国家在语料库研究领域已经形成比较专业的学术团队。2005 年至今，中国发表在 Web of Science 的论文总数 282 篇，占总数的 5.53%，但中心度为 0.01，这表明中国在语料库研究领域影响力很低，其研究水平需提高。

在 Cite Space 软件中将数据抽取阈值设置为 Top 50 perslice，可将世界各国发表的论文数量及时间以年轮的大小和颜色直观地展示。在得到的语料库领域研究的国家 / 地区综合分析知识图谱中（见图 5），共有 57 个结点，220 条连线，可以看出各国 / 地区间有较多合作，从而得出语料库研究地区大致可以分为 4 个中心，分别是美国、英国、德国和西班牙。一个节点的中介中心度越高，说明它在网络中最短路径上出现的越多，其影响力和重要程度越大 [17-18]。从节点中心度来看，美国的节点中心度最大，说明美国与其他语料库研究的地区存在某种程度上的合作关系，如英国、德国和西班牙等。从发文的突增性来看，南非的发文突增性最大，为 8.4，这说明南非在 2005-2017 年发表的与语料库主题相关的论文数量有较大的突破。

知识管理论坛，2018( 4 ):208–224
DOI：10.13266/j.issn.2095–5472.2018.021



表 6 语料库研究国家 / 地区文献发表情况

| 国家 / 地区 | 发文量 | 突现性 | 中心度 | 国家 / 地区 | 发文量 | 突现值 | 中心度 |
|---|---|---|---|---|---|---|---|
| USA | 895 | | 0.25 | NEW ZEALAND | 59 | | 0.03 |
| ENGLAND | 595 | | 0.24 | ISRAEL | 49 | | 0 |
| SPAIN | 573 | | 0.12 | BRAZIL | 42 | 4.43 | 0.02 |
| GERMANY | 487 | | 0.22 | IRAN | 41 | | 0 |
| BELGIUM | 338 | | 0.1 | SOUTH KOREA | 40 | | 0 |
| PEOPLES R CHINA | 282 | | 0.01 | CZECH REPUBLIC | 37 | | 0.02 |
| FRANCE | 240 | 7.3 | 0.03 | IRELAND | 32 | | 0.03 |
| NETHERLANDS | 158 | | 0.12 | DENMARK | 31 | | 0.01 |
| CANADA | 155 | | 0.04 | HUNGARY | 30 | 3.99 | 0 |
| AUSTRALIA | 154 | | 0.16 | RUSSIA | 30 | | 0.01 |
| ITALY | 149 | | 0.06 | ARGENTINA | 29 | | 0 |
| SOUTH AFRICA | 120 | 8.4 | 0.01 | PORTUGAL | 28 | | 0.01 |
| TAIWAN | 97 | | 0 | WALES | 27 | | 0.07 |
| CHILE | 89 | | 0 | TURKEY | 24 | | 0.02 |
| SWEDEN | 88 | | 0.03 | ESTONIA | 23 | 6.31 | 0 |
| JAPAN | 86 | | 0 | MALAYSIA | 19 | | 0.01 |
| SWITZERLAND | 86 | | 0.06 | SINGAPORE | 19 | | 0 |
| SCOTLAND | 83 | | 0.02 | SLOVENIA | 16 | | 0 |
| FINLAND | 71 | 3.12 | 0.02 | ROMANIA | 16 | 2.68 | 0 |
| NORWAY | 68 | | 0 | GREECE | 15 | | 0.03 |
| POLAND | 63 | | 0.01 | CROATIA | 13 | 3.3 | 0.01 |
| AUSTRIA | 60 | | 0.02 | MEXICO | 11 | | 0.02 |

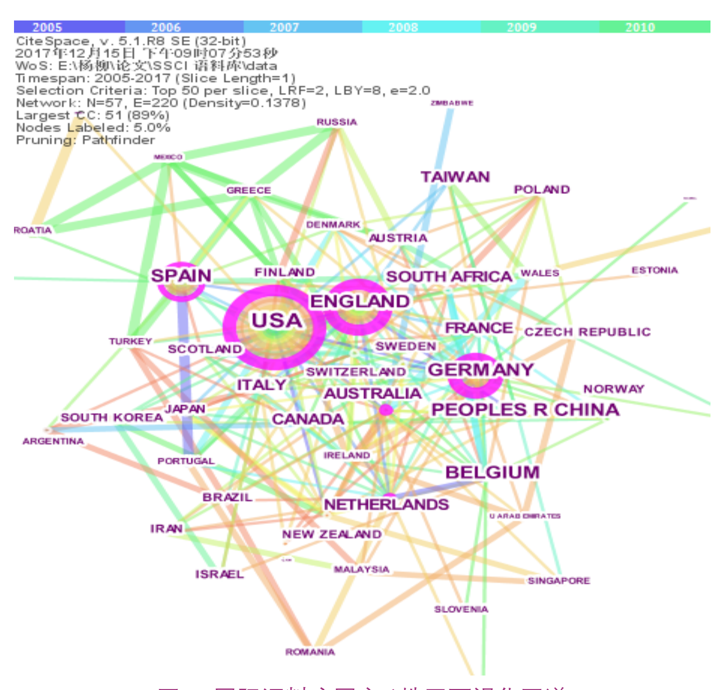



图 5 国际语料库国家 / 地区可视化图谱



知识管理论坛
2018 年 第 4 期（总第 16 期）

通过节点 Citation History 功能可以捕获中国这一节点两个方面的详细信息：一方面图 6 清晰展示了中国 2005-2017 的发文频次的变化情况，其中 2005-2017 近 13 年间中国在语料库领域的发文整体趋势上升；另一方面可以通过“Articles Published in This Country in 280 Records”的记录（即中国这一时期语料库的 282 篇施引文献的具体信息），进一步挖掘中国学者在语料库领域的分布信息。

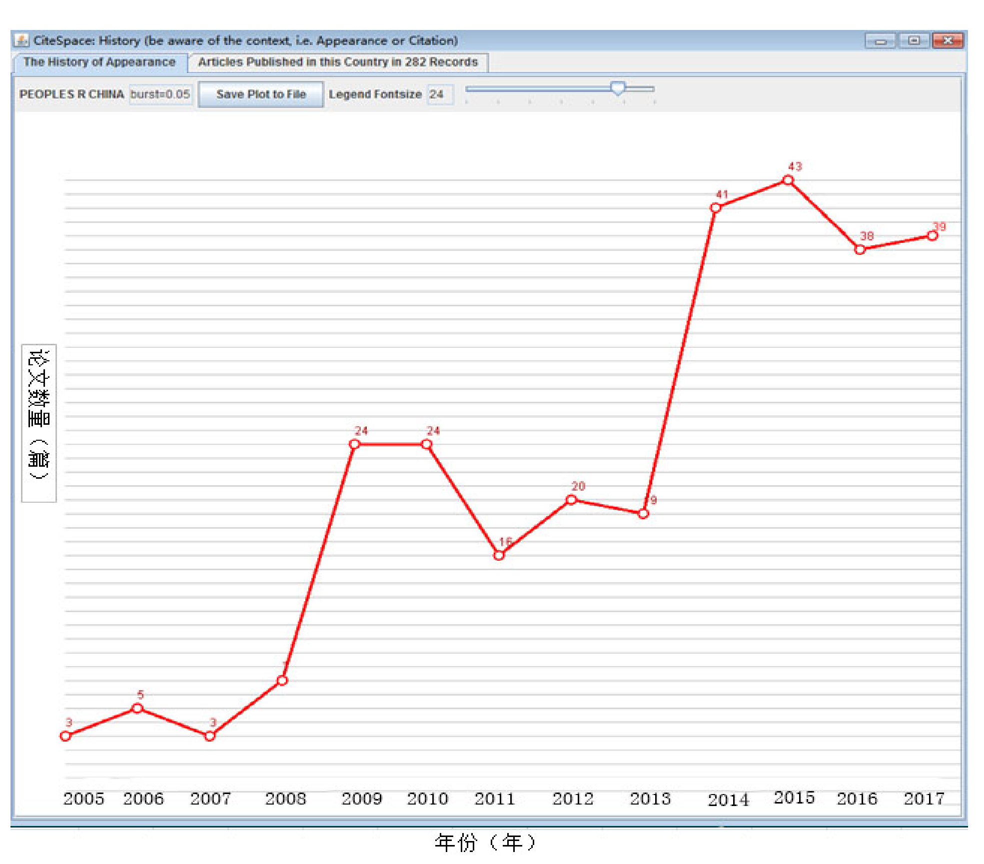



图 6 中国在语料库领域的发文历史



## 2.5 国际语料库研究机构分析

由于作者和机构之间具有从属关系，而机构在研究领域上具有不同的侧重点，期刊的机构分布不仅体现了该刊的侧重领域和方向，还体现了机构对刊物的支持和认同 [19]。因此，对研究机构进行分析，能够了解到某一领域核心科研机构的研究动态，掌握该领域的研究热点及发展趋势。因此，采用 CiteSpace 软件，将时区选择（Time Slicing）设定为 2005-2017 年，时间分区切片选择 1 年；节点类型（Node Type）选择机构（institution），修剪（Pruning）选择寻径算法（Pathfinder）和修剪切片网络（Pruning sliced network）[20]，运行 Citespace 软件，生成国际语料库研究机构的知识图谱，如图 7 所示：

图 7 中共有 184 个节点，102 条连线，网络密度为 0.006 1，这表明语料库研究仍处于发展阶段，大的成熟研究团体尚未形成，更广范围的机构合作有待形成。其中比较成熟的研究团队，如比利时根特大学为中心的连线较密，说明其与比利时鲁汶大学和比利时安特卫普大学有着较密切的合作。为了更清晰地了解国际科研机构对语料库领域的研究情况，统计语料库研究机构得到表 7，由表 7 可以看出，比利时根特大学、比利时鲁汶大学、英国兰卡斯

知识管理论坛，2018( 4 ):208–224
DOI：10.13266/j.issn.2095–5472.2018.021



特大学、英国爱丁堡大学、英国伯明翰大学、美国宾夕法尼亚州立大学、法国国家科学研究院、西班牙瓦伦西亚大学、比利时安特卫普大学、英国曼彻斯特大学和芬兰赫尔辛基大学等在语料库研究领域排在前 11 位， 这些学校在高等教育研究上处于领先的地位，其发文总数都在 30 篇以上，比利时根特大学发表 123 篇文献，文献数量排名第一。从中心度排序来看，比利时根特大学和比利时鲁汶大学的中心度最大，达到了 0.12，这说明这两种机构与其他机构合作广泛。从突现性来看，比利时安特卫普大学和西班牙瓦伦西亚大学突现性数值较大，这说明这两个机构在语料库研究上有较大的突破。

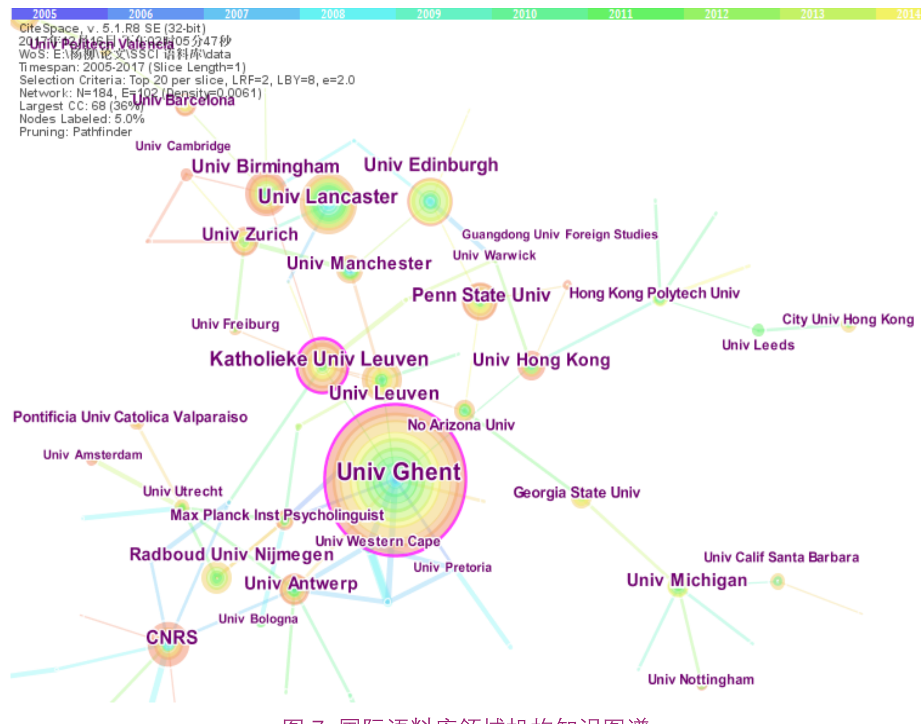



图 7 国际语料库领域机构知识图谱

表 7 Web of Science 收录语料库文献研究机构分布

| 机构 | 频次 | 中心度 | 突现值 |
|---|---|---|---|
| Ghent University ( 比利时根特大学 ) | 123 | 0.12 | |
| Katholieke Universiteit Leuven ( 比利时鲁汶大学 ) | 88 | 0.12 | 4.57 |
| Lancaster University ( 英国兰卡斯特大学 ) | 55 | 0 | |
| The University of Edinburgh ( 英国爱丁堡大学 ) | 45 | 0.01 | 4.17 |
| University of Birmingham ( 英国伯明翰大学 ) | 43 | 0.05 | |
| Penn State University ( 美国宾夕法尼亚州立大学 ) | 40 | 0.01 | |
| Centre National de la Recherche cientifique ( 法国国家科学研究院 ) | 38 | 0.07 | 4.77 |
| University of Valencia ( 西班牙瓦伦西亚大学 ) | 34 | 0 | 6.82 |
| University of Antwerp ( 比利时安特卫普大学 ) | 33 | 0.07 | 7.4 |
| The University of Manchester ( 英国曼彻斯特大学 ) | 32 | 0.01 | |
| University of Helsinki ( 芬兰赫尔辛基大学 ) | 30 | 0 | |

## 2.6 国际语料库被引文献分析

某一领域期刊论文被引用频次在一定程度上说明该领域的学术研究的理论水平和发展速度。通过分析这些论文，不仅可以直观地了解该研究领域在过去和当前的发展状况，还可以大概预测出其未来的发展趋势[21]。利用 CiteSpace 软件对文献数据进行可视化分析，网络节点（node types）为被引文献（cited reference），得到共被引文献图谱，如图 8 所示：

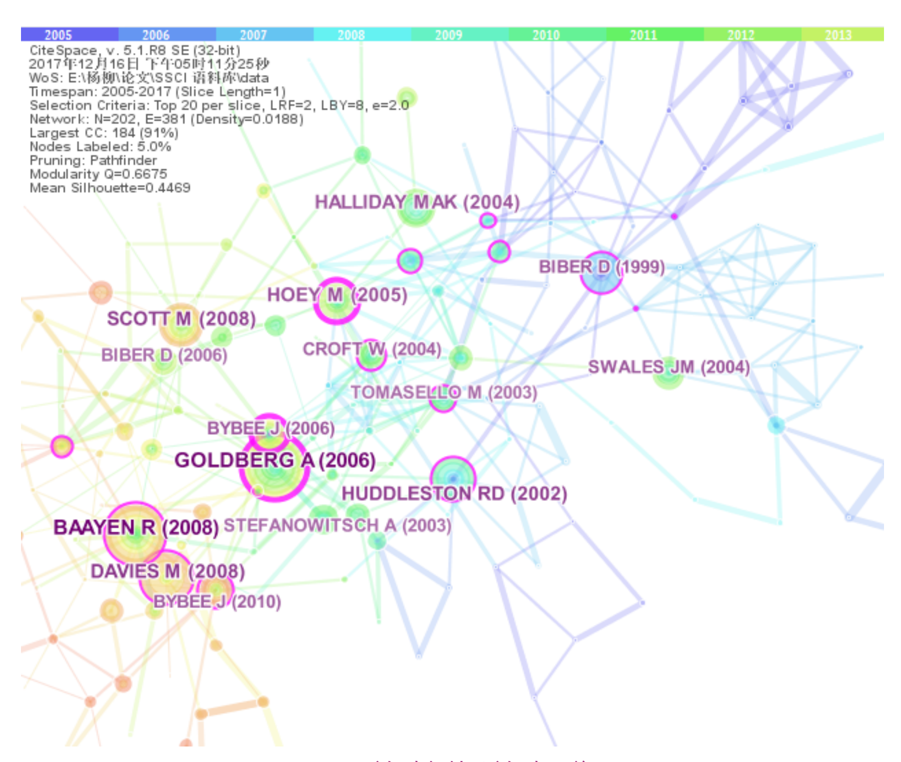



图 8 被引文献共被引图谱

从图 8 可以看出，此次分析共生成了 202 个节点，381 条连线，每个节点代表一篇被引文献，节点向外延伸的不同颜色与该文献所在年份的颜色相对应，节点越大表示被引频次越高，在一定程度上也代表该领域的研究重点。整理共被引网络图谱，其共被引频次排名前 5 和中心度大于 0.18 的数据见表 8 和表 9，美国学者 A. Goldberg 出版的著作 Constructions at Work 无论是被引频次还是中心度排名都在前三，这表明该书籍在语料库研究领域具高影响力。

表 8 共被引频次排名前 5 的被引文献

| 频次 | 文献名 | 作者（发表年份） |
| --- | --- | --- |
| 78 | *Analyzing linguistic data: a practical introduction to statistics using R* | R. Baayen（2008） |
| 77 | *Constructions at Work* | A. Goldberg (2006) |
| 59 | *The corpus of contemporary American English--a useful tool for English teaching and research* | M. Davies（2008） |
| 57 | *Wordsmith tools version 5, Liverpool: Lexical Analysis Software Ltd* | M. Scott（2008) |
| 56 | *The Cambridge grammar of the English language* | R. D. Huddleston（2002） |

知识管理论坛，2018( 4 ):208–224
DOI：10.13266/j.issn.2095–5472.2018.021

表 9  共被中心度大于 0.18 的被引文献

| 频次 | 文献名 | 作者（发表年份） |
|---|---|---|
| 0.28 | *A usage-based exemplar model approach to Spanish verbs of "becoming"* | J. Bybee (2006) |
| 0.25 | *Lexical priming: a new theory of words and language* | M. Hoey（2005） |
| 0.21 | *Constructions at work* | A. Goldberg (2006) |
| 0.19 | *Constructing a Language: a usage-based theory of language acquisition* | M. Tomasello (2003) |
| 0.18 | *Language, usage and cognition* | J. Bybee（2010） |
| 0.18 | *An academic formulas list (AFL)* | R. Simpson-Vlach（2010） |
| 0.18 | *Humble servants of the discipline? self-mention in research article* | K. Hyland（2001） |

在 5 篇共被引频次最高的文献中有 4 本专著都带有工具书性质，Analyzing linguistic data: a practical introduction to statistics using R 是面向非数学背景学者展示怎样用 R 语言进行语言学语料分析；The corpus of contemporary American English--a useful tool for English teaching and research 介绍美国当代英语语料库（COCA）在英语教学和研究中的应用；Wordsmith tools version 5, Liverpool: Lexical Analysis Software Ltd 介绍语料库工具 Wordsmith；The Cambridge grammar of the English language 是基于描写语法的英语辞书，例句均来自真实语料。这些高频次共被引文献揭示出语料库语言学的一个重要特征，即实践性和工具性。语料库语言学是理论与实践的结合，尤其实践性是其突出特点。并且语料库工具和语料库本身都在不断更新、升级，Wordsmith 现在已经更新至 7.0 版本，COCA 已经有了库容达到 1.4 亿的升级版 iWeb。

在共被引最高频次和中心度最高的文献中均入选的 *Constructions at work: the nature of generalization in language*（《运作中的构式：语言概括性的本质》）[21] 具有重要的理论价值，是构式语法的扛鼎之作。其作者 A. Goldberg 提出的“构式”概念引起了整个语言学界的极大关注，其发展势头极为迅猛。某种意义上，构式主义已经形成独立的研究流派。语料库与构式的结合既出于理论上的契合，又凸显了语料库的工具优势。

### 2.7 国际语料库研究前沿热点

关键词是一篇文献的核心与精髓，是对主题的概括与凝练，反映文献的核心内容，也是文献计量研究的重要指标，当多篇文章的关键词具有一致性时，这些文章的主题或多或少在一定程度上具有相关性。关键词共现知识图谱能够将具有相同关键词的文章进行聚类，进而体现出同一研究领域的关键节点，集中展现一段时间内相关文献的研究热点，有利于从整体上把握已有研究内容。同时，通过对关键词共现产生的中心性分析可以揭示出研究热点之间的转化关系，因此，本文利用已收集的文献数据库的关键词来分析语料库的研究热点。在 CiteSpace 软件中，将节点类型设置为“Keyword”，对 5 096 篇文献进行关键词共现分析得到关键词共现的研究热点图谱，运行结果表明，共有 323 个节点，930 条连线，且密度为 0.017 9，如图 9 所示。图 9 中带有紫色圆圈的关键词具有高中心性，是一个研究热点向另一个研究热点转化的重要转折点。

通常频次高的关键词被用来确定一个研究领域的热点，表 10 列出了共现频次大于等于 30 的关键词及其序号、频次、突现值和中心度。从表 10 中可以看出，语料库语言学（corpus linguistics）作为关键词共现频次最多，有 238 次，且中心度为 0.14，处于第 5 位，其中西班牙语（Spanish）、话语分析（discourse analysis）、语料库（corpora）、词典编纂（lexicography）和句法（syntax）的突现值均非常高，表明这 5 个关键词是各自年份的热点。







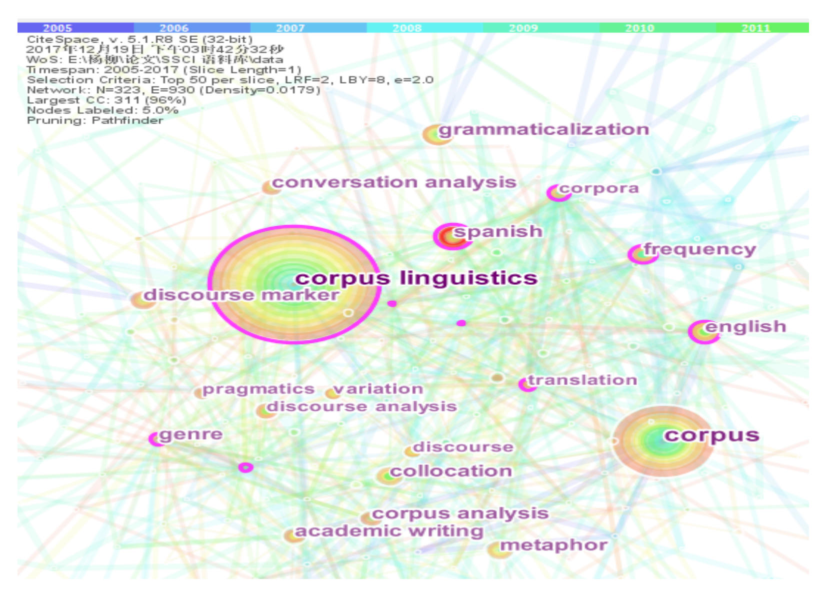



图 9 关键词分析可视化图谱

表 10 关键词频次和中心度

| 序号 | 关键词 | 频次 | 突现值 | 中性度 |
|---|---|---|---|---|
| 1 | corpus linguistics | 238 | | 0.14 |
| 2 | corpus | 171 | | 0.09 |
| 3 | Spanish | 79 | 3.14 | 0.1 |
| 4 | grammaticalization | 78 | | 0.09 |
| 5 | collocation | 71 | | 0.06 |
| 6 | corpus analysis | 69 | | 0.08 |
| 7 | frequency | 69 | | 0.16 |
| 8 | English | 67 | | 0.17 |
| 9 | metaphor | 66 | | 0.06 |
| 10 | conversation analysis | 65 | | 0.1 |
| 11 | discourse marker | 62 | | 0.07 |
| 12 | academic writing | 58 | | 0.07 |
| 13 | genre | 57 | | 0.14 |
| 14 | variation | 52 | | 0.05 |
| 15 | discourse analysis | 50 | 3.17 | 0.04 |
| 16 | discourse | 50 | | 0.08 |
| 17 | translation | 50 | | 0.11 |
| 18 | pragmatics | 50 | | 0.06 |
| 19 | corpora | 49 | 2.69 | 0.15 |
| 20 | gender | 45 | | 0.16 |
| 21 | prosody | 45 | | 0.07 |
| 22 | French | 39 | | 0.1 |
| 23 | syntax | 36 | 3.51 | 0.05 |
| 24 | Dutch | 35 | | 0.03 |
| 25 | evaluation | 31 | | 0.07 |
| 26 | lexicography | 31 | 9.12 | 0.05 |
| 27 | semantics | 30 | | 0.12 |

**Top 22 Keywords with the Strongest Citation Bursts**

| Keywords | Year | Strength | Begin | End | 2005 - 2017 |
|---|---|---|---|---|---|
| representation | 2005 | 5.4444 | 2005 | 2007 | |
| conversation analysis | 2005 | 4.5006 | 2005 | 2009 | |
| metadiscourse | 2005 | 3.9206 | 2005 | 2008 | |
| language production | 2005 | 3.6259 | 2005 | 2007 | |
| lexicography | 2005 | 12.0838 | 2006 | 2009 | |
| dictionary | 2005 | 5.0203 | 2006 | 2007 | |
| phonology | 2005 | 2.9652 | 2006 | 2007 | |
| complexity | 2005 | 4.8073 | 2007 | 2010 | |
| politeness | 2005 | 9.9279 | 2008 | 2010 | |
| model | 2005 | 2.9226 | 2008 | 2012 | |
| metaphor | 2005 | 2.908 | 2008 | 2011 | |
| gender | 2005 | 7.9156 | 2009 | 2010 | |
| comprehension | 2005 | 3.4039 | 2009 | 2010 | |
| morphology | 2005 | 8.2168 | 2010 | 2011 | |
| syntax | 2005 | 7.309 | 2010 | 2013 | |
| identity | 2005 | 7.8008 | 2012 | 2013 | |
| genre | 2005 | 7.5246 | 2012 | 2013 | |
| word | 2005 | 7.3909 | 2012 | 2013 | |
| discourse marker | 2005 | 4.863 | 2014 | 2017 | |
| collocation | 2005 | 4.2946 | 2014 | 2015 | |
| corpus analysis | 2005 | 10.6228 | 2015 | 2017 | |
| german | 2005 | 8.7786 | 2015 | 2017 | |

图 10 高突变关键词时间线图谱

图 10 清晰地显示了 2005-2017 年研究热点关键词的演变，但也需要具体甄别和阐释，如单独看“model”和“corpus analysis”没有意义。整体看语料库 2005-2017 年的研究热点包括话语分析(conversation analysis、metadiscourse、politeness、discourse marker)、词典编撰(lexicography、dictionary)、词汇(morphology、word、collocation)，此外还

知识管理论坛，2018( 4 ):208-224
DOI：10.13266/j.issn.2095-5472.2018.021

有音系学（phonology）、句法（syntax）、隐喻( metaphor )、性别( gender )、身份( identity )、文类（genre）。以上分析显示一方面词典编撰和词汇搭配一直都是语料库语言学的重要课题，另一方面隐喻、性别、身份和文类等关键词往往和话语分析、文学研究及专门用途语言相关，这表明语料库语言学的研究正在拓展到更多领域，并更加细致。延续到 2017 年的关键词包括话语标记、搭配（collocation）和德语（German）。德语成为一个热点关键词可能有两个原因：①有 4 本德国出版的期刊均为语料库核心期刊，研究成果发表渠道较为丰富；②近年来关于德语的研究比较活跃，如古 / 中高地德语的语料库建设和研究等。

## 3 结论

从文献发表数量看，语料库语言学研究经历了快速发展，近年来每年均有大量高质量成果发表。在 WOS 数据库共收录期刊文献 5 096 篇，年均文献量为 392 篇，分布在 251 个刊物，其中核心期刊有 19 种，欧洲国家出版刊物占 15 本，美洲 3 本，非洲 1 本。这些期刊文献反映了近 13 年的语料库语言学研究的最高水平，可重点关注。此外，荷兰、英国、德国均拥有 4 本及以上核心期刊，形成语料库研究的中心，反过来进一步促进了本国研究的发展，比如德语是近 3 年来的持续热点。事实上，针对印欧语系语言的研究的确在语料库研究中占据主流，针对其他语言的研究一方面极具必要性，另一方面在发表渠道上不占优势。目前我国北京外国语大学和上海交通大学一北一南形成语料库研究的两个核心，在创建英文期刊、进入国际学界方面大有可为。

在 3 755 位第一作者中，S. T. Gries、G. M. de Schryver、K. Hyland、G. Parodi 和 D. Biber 等拥有最高发文量。被引文献共涉及作者 85 996 人，其中高被引作者共 43 人，D. Biber、K. Hyland、J. Bybee、M. A. K. Halliday、E. C. Traugott、W. Labov 和 S. T. Gries 位居前列，这些作者在国际语料库领域都做出了卓越贡献。分析作者合作水平后发现，语料库研究的合作度为 0.82，合作率为 22.36%，合作度和合作率都很低。共被引文献还揭示出语料库语言学的一个突出特点是兼具理论性与实践性。高被引文献的作者往往也是重要的语料库建设者及软件开发者。此外，语料库的工具性并不能掩盖其理论价值，对理性主义的矫正，和构式语法的结合，语料库的建立对语言习得、翻译和语言本质的认识均有重要影响。

共有 57 个国家对语料库进行了研究，美国、英国、西班牙、德国、比利时、中国和法国等是开展语料库研究的主要国家，比利时根特大学、比利时鲁汶大学、英国兰卡斯特大学、英国爱丁堡大学、英国伯明翰大学、美国宾夕法尼亚州立大学、法国国家科学研究院、西班牙瓦伦西亚大学、比利时安特卫普大学、英国曼彻斯特大学和芬兰赫尔辛基大学等在语料库研究领域排在前 11 位，处于领先的地位。中国学者在进行访问交流时可重点考虑这些学校。

词频、搭配、词典编撰与语料库语言学具有天然的联系，一直是重要的研究内容，词频和搭配也是展开其他研究的重要手段。近些年来，语料库研究越来越呈现出跨学科、多角度的特点。专门用途语言、学术英语，不同文类、不同语种均成为研究对象。性别、身份、隐喻及话语分析成为新的研究热点。基于语料库的话语分析、语料库文体学相继涌现，与构式语法、认知语言学的结合是研究前沿。我国在语料库翻译学、学习者语料库、汉语语料库建设方面成果颇丰，是国际语料库语言学研究的一部分。对国际研究热点和前沿的关注有利于人们拓展和深入现有研究，也有利于与国际学界进行更有效的对话。

参考文献：

[1] QUIRK R. Words at work: lectures on textual structure [M]. Singapore: NUS Press, 1986.
[2] LEECH G. Corpora, the linguistics encyclopedia[M]. London:Routledge, 1991.

[3] McENERY T, XIAO R, TONO Y. Corpus-based language studies: an advanced resource book[M]. London: Routledge, 2006.
[4] SINCLAIR J. Corpus, concordance, collocation[M]. Oxford: Oxford University Press, 1991.
[5] BAKER M. Corpus linguistics and translation studies: implications and applications [C]// BAKER M, FRANCIS G, TOGNINI-BONELLI E. Text and technology: in honour of John Sinclair. Philadelphia: John Benjamins, 1993: 233-250.
[6] ATKINS S, CLEAR J, OSTLER N. Corpus design criteria[J]. Literary and linguistic computing, 1992, 7(1): 1-16.
[7] RENOUF A. Teaching corpus linguistics to teachers of English [C]//WICHMAN A, FLIGELSTONE S, McENERY T, et al. Teaching and language corpora. New York: Longman, 1997: 255-266.
[8] 桂诗春 , 杨惠中 . 中国学习者英语语料库 [M]. 上海 : 上海外语教育出版社 , 2003.
[9] 王克非 . 语料库翻译学——新研究范式 [J]. 中国外语 , 2006(3): 8-9.
[10] 崔雷 , 刘伟 , 闫雷 . 文献数据库中书目信息共现挖掘系统的开发 [J]. 现代图书情报技术 , 2008(8): 70-75.
[11] 杨利军 , 吴智君 . 低被引文献对布拉德福定律的影响研究 [J]. 情报理论与实践 , 2016, 39(9): 43-46.
[12] 褚旭 , 熊华军 . 2000 年以来我国教育技术论文作者可视化分析——基于《中国电化教育》和《电化教育研究》载文 [J]. 重庆高教研究 , 2015(6): 100-108.
[13] 洪波 . 我国高等职业教育研究的知识图谱分析——基于 1992-2016 年核心期刊文献 [J]. 职业技术教育 , 2017, 38(6): 45-50.
[14] 孙雨生 , 陈卫 . 我国网格服务研究进展——基于 CNKI(2003-2012) 的文献计量与知识图谱分析 [J]. 现代情报 , 2013, 33(7): 102-111.
[15] 廉同辉 , 余菜花 , 宗乾进 . 我国旅游网站的网络结构研究——基于社会网络分析法 [J]. 旅游科学 , 2012, 26(6): 80-88.
[16] WALTMAN L, van ECK N J, van LEEUWEN T N, et al. Towards a newcrown indicator: an empirical analysis[J]. Scientometrics, 2011, 87(3): 467-481.
[17] 刘则渊 , 陈悦 , 侯海燕 , 等 . 科学知识图谱 : 方法与应用 [M]. 北京 : 人民出版社 , 2008.
[18] 林德明 , 陈超美 , 刘则渊 , 等 . 共被引网络中介中心性的 Zipf-Pareto 分布研究 [J]. 情报学报 , 2011, 30(1): 76-82.
[19] 姜春林 , 胡志刚 . 《管理学报》2004-2009 年载文计量分析 [J]. 管理学报 , 2010, 7(8): 1137-1143.
[20] 索璠冰 . 基于 CiteSpace 和文献计量的国内云计算研究现状分析 [J]. 图书情报导刊 , 2017, 2(6): 60-65.
[21] 李旭辉 , 李超 , 魏瑞斌 , 等 . 基于 CSSCI 的信息消费被引文献计量研究 [J]. 图书馆工作与研究 , 2014(4): 104-108.
[22] GOLDBERG A. Constructions at work: the nature of generalization in language[M]. Oxford: Oxford University Press, 2006.

## Information Visualization Analysis on the Research Hot Spots and Frontiers of International Corpus Linguistics

Yang Liu
School of English Studies, Shanghai International Studies University, Shanghai 200083

**Abstract: [Purpose/significance]** This paper aims at grasping the overall context and research findings of international corpus research in recent years, clarifying the hot spots and exploring the research frontiers in this field. **[Method/process]** It took the research papers published between 2005 and 2017 in Web of Science as data source, and made calculate analysis and knowledge domains map on these data through the softwares including BICOMB, Ucinet6 and CiteSpace from the following aspects: publication numbers and chronological distribution, source journals, research subjects, national and regional distribution, research institutions and cited literature. **[Result/conclusion]** It found that corpus study presents the characteristics of interdisciplinary and multi-angle, and it's research objects involve more languages and literature genre. Collocation and lexicography keep being important studies while special purpose language, academic English, gender, identity, metaphor and discourse analysis turn into new research hotspots, and the combination with construction grammar and cognitive linguistics are the research frontiers.

**Keywords:** corpus visualized analysis BICOMB Ucinet6 CiteSpace